to be encoded by the human genome. Because of its tremendous size, to date only a portion of the human genome has been sequenced and deposited in genome sequence databases, and the positions of many genes and their exact nucleotide sequences remain unknown. Moreover, the biological function(s) of the gene products encoded by many of the genes sequenced so far remain unknown. Similar situations exist with respect to the genomes of many other organisms.

Notwithstanding such complexities, numerous genome sequence efforts designed to determine the exact sequence of the nucleotides found in genomic DNA of various organisms are underway and significant progress has been made. For example, the Human Genome Project began with the specific goal of obtaining the complete sequence of the human genome and determining the biochemical function(s) of each gene. To date, the project has resulted in sequencing a substantial portion of the human genome (J. Roach, http://weber.u.washington.edu/~roach/human_genome_progress2.html) (Gibbs, 1995), and is on track for its scheduled completion in the near future. At least twenty-one other genomes have already been sequenced, including, for example, *M. genitalium* (Fraser *et al.*, 1995), *M. jannaschii* (Bult *et al.*, 1996), *H. influenzae* (Fleischmann *et al.*, 1995), *E. coli* (Blattner *et al.*, 1997), and yeast (*S. cerevisiae*) (Mewes *et al.*, 1997). Significant progress has also been made in sequencing the genomes of model organisms, such as mouse, *C. elegans*, and *D. melanogaster*. Several databases containing genomic information annotated with some functional information are maintained by different organizations, and are accessible via the internet, for example, http:/www.tigr.org/tdb; http://www.genetics.wisc.edu; http://genome-www.stanford.edu/~ball; http://hiv-web.lanl.gov; http://www.ncbi.nlm.nih.gov; http://www.ebi.ac.uk; http://pasteur.fr/other/biology; and, http://www-genome.wi.mit.edu.

Such sequencing projects result in vast amounts of nucleotide sequence information, which is typically deposited in genome sequence databases. However, these raw data (much of it being known only at the cDNA level), being devoid of

corresponding information about genes and protein structure or function, are in and of themselves of extremely limited use (Koonin, *et al.* (1998), Curr. Opin. Struct. Biol., vol. 8:355-363). Thus, the practical exploitation of the vast numbers of sequences in such genome sequence databases is crucially dependent on the ability to identify genes and, for example, the function(s) of gene-encoded proteins.

To maximize the utility of such nucleotide sequence information, it must be interpreted. Various tools have been developed to assist in this process. For example, algorithms have been developed to analyze what a particular nucleotide sequence encodes, *e.g.*, a regulatory region, an open reading frame (ORF), particularly for protein sequences, or a non-translated RNA, based on homology with known sequences (which are presumed to have similar structures and related functions). *See, e.g.*, "Frames" (Genetics Computer Group, Madison, WI; www.gcg.com), which is used for identifying ORFs. For sequences predicted or determined to be ORFs, it is possible to determine the amino acid sequence of the protein encoded thereby using simple analytical tools well known in the art. For example, *see* "Translate" (Genetics Computer Group, Madison, WI; www.gcg.com). However, to date determination of the primary structure of a protein in and of itself provides little, if any, functional information about the protein or its corresponding gene. Thus, the ability to predict the three-dimensional structure of a protein from its amino acid sequence is of great theoretical[1,2] and practical importance.[3]

In practice, structure prediction can be attempted on various levels, ranging from purely *de novo*, or "*ab initio*," approaches to those that incorporate constraints derived from experimental data. The latter aspect of protein structure modeling has recently attracted significant attention[4-6] due to its possible application to model building based on structural constraints provided by nuclear magnetic resonance (NMR),[7] x-ray crystallography, or other experimental methods.

Perhaps the most useful method developed to date for predicting three-dimensional protein structures is the MONSSTER (Modeling of New Structures from Secondary and Tertiary Restraints) algorithm.[8] MONSSTER provides a well-

defined protocol for identifying moderate-resolution native-like three-dimensional structures from known secondary structure and a small number of tertiary constraints based on alpha-carbon ("Cα") positions the amino acid residues of the protein.

That having been said, when a large number of distance constraints between atoms comprising amino acid residues of a protein are obtained from NMR or other experimental methods, and possibly from homology-based theoretical models of protein structure, more standard algorithms[9-12] are the tools of choice. These algorithms are based on purely geometrical considerations, followed by restrained molecular dynamics refinement of the model structures.[13] However, in many real life situations, the number of available geometric constraints (*e.g.*, interatomic distances, bond angles, *etc.*) is relatively small and limited, particularly in the early stages of structure determination based on experimental methods such as NMR. When the available geometric constraints are too sparse to define even a moderate resolution structure (*i.e.*, a cRMSD of about 4-6 Å), it is necessary to use modeling methods that employ a reasonable force field capable of providing an overall protein-like bias. In such a case, even a small number of distance constraints could be sufficient to guide folding to the correct structure. Due to the necessity of sampling a substantial part of the protein conformational space, any such an algorithm should be computationally efficient. Moreover, the force field of the model should be able to correct for. The MONSSTER method is relatively efficient from a computational standpoint and can compensate for some errors in the provided set of Cα distance constraints. Significantly, however, even though MONSSTER is relatively efficient, the computational demands of that method have limited its application to proteins containing no more than 150 amino acid residues.

In the past several years, there also have been a number of other studies that have utilized experimentally derived secondary structure and a limited number of known, experimentally derived tertiary constraints to predict the global fold of a globular protein. In particular, Smith-Brown *et al.*[4] reported the modeling of a protein as a chain of glycine residues. Tertiary constraints were reported to have